机器学习导论
习题四

151250104, 卢以宁, kiwiloveskiwis@gmail.com

2017 年 5 月 17 日

1 [20pts] Reading Materials on CNN

卷积神经网络 (Convolution Neural Network, 简称 CNN) 是一类具有特殊结构的神经网络, 在深度学习的发展中具有里程碑式的意义。其中, Hinton 于 2012 年提出的AlexNet可以说是深度神经网络在计算机视觉问题上一次重大的突破。

关于 AlexNet 的具体技术细节总结在经典文章"ImageNet Classification with Deep Convolutional Neural Networks", by Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton in NIPS'12, 目前已逾万次引用。在这篇文章中, 它提出使用 ReLU 作为激活函数, 并创新性地使用 GPU 对运算进行加速。请仔细阅读该论文, 并回答下列问题 (请用 1-2 句话简要回答每个小问题, 中英文均可)。

(a) **[5pts]** Describe your understanding of how ReLU helps its success? And, how do the GPUs help out?

(b) **[5pts]** Using the average of predictions from several networks help reduce the error rates. Why?

(c) **[5pts]** Where is the dropout technique applied? How does it help? And what is the cost of using dropout?

(d) **[5pts]** How many parameters are there in AlexNet? Why the dataset size(1.2 million) is important for the success of AlexNet?

关于 CNN, 推荐阅读一份非常优秀的学习材料, 由南京大学计算机系吴建鑫教授[1]所编写的讲义 Introduction to Convolutional Neural Networks[2], 本题目为此讲义的 Exercise-5, 已获得吴建鑫老师授权使用。

Solution. 此处用于写解答 (中英文均可)

(a) **[5pts]** ReLU is desirable for its ability of preventing saturating. Other saturating non-linear functions, such as sigmoid, will lead the magnitude of the gradient to significantly

[1] 吴建鑫教授主页链接为cs.nju.edu.cn/wujx

[2] 由此链接可访问讲义https://cs.nju.edu.cn/wujx/paper/CNN.pdf

reduce or even vanish[3]. Due to the same reason, the training of ReLU would be faster. AlexNet employs two GPUs which are enough to contain a big network, while they are also welled-suited to cross-GPU communication. Besides, it could be more convenient to ”tune the amount of communication”

(b) **[5pts]** It fixes high-variance and prevents overfitting.

(c) **[5pts]** Dropout technique is used in ”the first two fully-connected layers”, where ”the output of each hidden neuron is set to zero with probability 0.5”. In this case the features learned can be more robust, therefore ”useful in conjunction with many different random subsets of other neurons”. The cost of dropout is the doubled number of iterations.

(d) **[5pts]** 60 million parameters. A large dataset can help reduce variance and relieve overfittting.

2 [20pts] Kernel Functions

(1) 试通过定义证明以下函数都是一个合法的核函数：

(i) **[5pts]** 多项式核: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j)^d$;

(ii) **[10pts]** 高斯核：$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$, 其中 $\sigma > 0$.

(2) **[5pts]** 试证明 $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{x}_j}}$ 不是合法的核函数。

Proof. (1) (i) Polynomial: First, K is symmetric.

Second, For a linear kernel where $d = 1$, $\mathbf{z^T K z} = \mathbf{z}(\mathbf{X^T X})\mathbf{z} = (\mathbf{Xz})^{\mathbf{T}}(\mathbf{Xz}) = \|\mathbf{Xz}\|^{\mathbf{2}} \geq 0$ Therefore the linear kernel is positive semi-definite(psd). Since the Hadamard product[4] of two psd matrices are as well psd, the polynomial kernel is valid.

(ii) Gaussian: First, K is symmetric.

Second, Let Y be a random variable such that $Y \sim N(\mu, \sigma^2)$.

Then $E(e^{aY}) = e^{\frac{a^2\sigma^2}{2} + a\mu}$, when$Y \sim N(0,1)$, $E(e^{ibY}) = e^{-\frac{b^2}{2}}$.

Replace b with $\frac{x_i - x_j}{\sigma}$ we have $\exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) = E(e^{iY(x_i - x_j)/\sigma})$

$\mathbf{z^T K z} = \sum_{j,k=1}^{n} z_j\, z_k\, h(x_j - x_k) = \sum_{j,k=1}^{n} z_j\, z_k \,\mathrm{E}\left[e^{i(x_j - x_k)Y/\sigma}\right]$

$= \mathrm{E}\left[\sum_{j,k=1}^{n} z_j\, e^{ix_jY/\sigma}\, z_k\, e^{-ix_kY/\sigma}\right] = \mathrm{E}\left[\left|\sum_{j=1}^{n} z_j\, e^{ix_jY/\sigma}\right|^2\right] \geq 0$ [5]

(2) Suppose that $n = 2, x_1 = (0,1), x_2 = (0,2)$, then $K = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \kappa(\mathbf{x}_1, \mathbf{x}_2) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \kappa(\mathbf{x}_2, \mathbf{x}_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-1}} & \frac{1}{1+e^{-2}} \\ \frac{1}{1+e^{-2}} & \frac{1}{1+e^{-4}} \end{bmatrix}$, where $\det(K) < 0$. Therefore it’s not a valid kernel.

□

[3] The CNN handout by Prof. Wu

[4] Hadamard product: https://www.wikiwand.com/en/Hadamard_product_(matrices)

[5] Other proof: https://stats.stackexchange.com/questions/35634/how-to-prove-that-the-radial-basis-function-is-a-kernel

3 [25pts] SVM with Weighted Penalty

考虑标准的 SVM 优化问题如下 (即课本公式 (6.35)),

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{3.1}$$

注意到，在(**??**)中，对于正例和负例，其在目标函数中分类错误的“惩罚”是相同的。在实际场景中，很多时候正例和负例错分的“惩罚”代价是不同的，比如考虑癌症诊断，将一个确实患有癌症的人误分类为健康人，以及将健康人误分类为患有癌症，产生的错误影响以及代价不应该认为是等同的。

现在，我们希望对负例分类错误的样本 (即 false positive) 施加 $k > 0$ 倍于正例中被分错的样本的“惩罚”。对于此类场景下，

(1) **[10pts]** 请给出相应的 SVM 优化问题;

(2) **[15pts]** 请给出相应的对偶问题，要求详细的推导步骤，尤其是如 KKT 条件等。

Solution. (1)

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C(\sum_{i\in pos}^{m}\xi_i + k\cdot\sum_{i\in neg}^{m}\xi_i) \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m \end{aligned} \tag{3.2}$$

(2) Let α, μ denote the Lagrange multipliers.

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \mu) = & \frac{1}{2}\|\mathbf{w}\|^2 + C(\sum_{i\in pos}^{m}\xi_i + k\cdot\sum_{i\in neg}^{m}\xi_i) \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{3.3}$$

Let $\nabla_{\mathbf{w}}L = \nabla_b L = \nabla_{\xi_i}L = 0$, we have:

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \\ 0 &= \sum_{i=1}^{m}\alpha_i y_i \\ C &= (\alpha_i + \mu_i)\cdot((i \in pos)?1 : \frac{1}{k}) \end{aligned} \tag{3.4}$$

Then the dual problem is :

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right) \\ \text{s.t.} \quad & \sum_{i=1}^{m} y_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq C\cdot((i \in pos)?1 : k) \end{aligned} \tag{3.5}$$

The KKT conditions are:

$$\begin{cases} \alpha_i, \mu_i, \xi_i \geq 0 \\ \xi_i - 1 + y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 0 \\ \alpha_i(1 - \xi_i - y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)) = 0 \\ \mu_i\xi_i = 0 \end{cases}$$

4 [35pts] SVM in Practice - LIBSVM

支持向量机 (Support Vector Machine，简称 SVM) 是在工程和科研都非常常用的分类学习算法。有非常成熟的软件包实现了不同形式 SVM 的高效求解，这里比较著名且常用的如 LIBSVM[6]。

(1) **[20pts]** 调用库进行 SVM 的训练，但是用你自己编写的预测函数作出预测。

(2) **[10pts]** 借助我们提供的可视化代码，简要了解绘图工具的使用，通过可视化增进对 SVM 各项参数的理解。详细编程题指南请参见链接：http://lamda.nju.edu.cn/ml2017/PS4/ML4_programming.html.

(3) **[5pts]** 在完成上述实践任务之后，你对 SVM 及核函数技巧有什么新的认识吗？请简要谈谈。

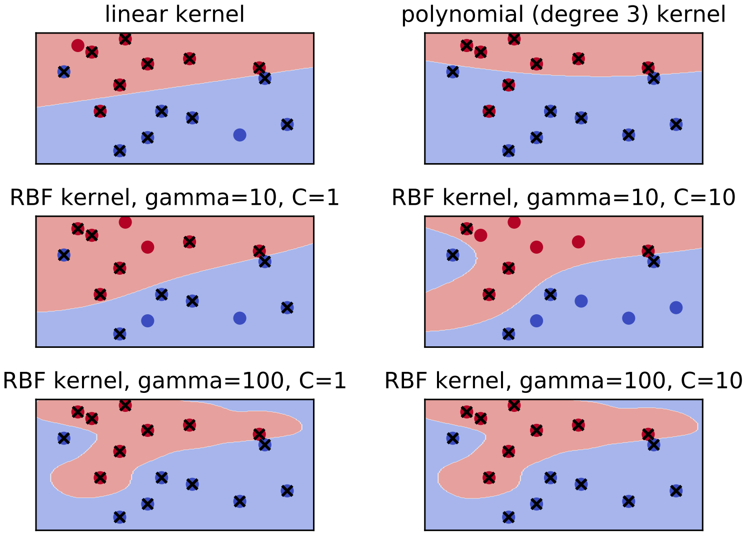

图 1: Support Vectors

Solution. The support vectors are marked with "X".
Through observation, the C parameter determines "the relative importance between large margin and small total price($\sum \xi_i$)", which invokes the trade-off between training-error penalty and stability. A large C provides low bias and high variance and vice versa. Another

[6] LIBSVM 主页课参见链接：https://www.csie.ntu.edu.tw/~cjlin/libsvm/

parameter $\gamma \propto \frac{1}{\sigma^2}$ intuitively implies the influence range of a single support vector. (i.e. large γ → small variance → the support vector does not have wide-spread influence.") According to the documentation, lower γ leads to smoother models and large γ may lead to overfitting. [7] When γ and C are small, the RBF kernel is able to simulate linear kernel.

[7] Reference: http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html#sphx-glr-auto-examples-svm-plot-rbf-parameters-py